DOCUMENT RESUME

ED 314 876 EA 021 651

TITLE	Gauging Student Performance. A Position Paper.
INSTITUTION	New York State School Boards Association, Albany.
PUB DATE	89
NOTE	42p.; Black and white photographs will not reproduce well.
PUB TYPE	Guides - Non-Classroom Use (055) -- Viewpoints (120)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Boards of Education; Elementary Secondary Education; *Standardized Tests; Testing; *Testing Problems; Test Interpretation; *Test Results; Test Validity
IDENTIFIERS	*New York

ABSTRACT

Testing and assessment have become extremely important components of the educational program. The purpose of this position paper is to provide factual and policy-related information that may be of use to school boards as they design or improve their districts' testing and assessment policies. It is the New York State policy that school boards should have a written testing policy. Current issues surrounding the use of standardized tests are as follows: (1) the impact on instruction of overreliance on standardized testing; (2) the psychological and social implications of overreliance on standardized testing; (3) misinterpretation of test results; (4) improving tests and testing; and (5) testing and assessment at the state and national levels. The booklet is prefaced by a summary of recommendations, and a sample testing program policy statement is appended. (41 references) (SI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314876

EA 021 651

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jeffrey M. Bowen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Board of Directors

President	JUDITH H. KATZ *Williamsville/Erie 1 BOCES*
Vice President	ALBERT W. HAWK *Livingston-Steuben-Wyoming BOCES*
Vice President	GORDON S. PURRINGTON *Guilderland*
Treasurer	EARL LUMLEY *St. Lawrence-Lewis BOCES*
Immediate Past President	REXENE ASHFORD-ORNAUER
Area 1	ALISON HYDE *East Aurora*
Area 2	ALBERT W. HAWK *Livingston-Steuben-Wyoming BOCES*
Area 3	PATRICIA M. PETESEN *Pioneer*
Area 4	DANIEL F. SCHULTZ *Skaneateles*
Area 5	KENNETH J. KAZANJIAN *New Hartford*
Area 6	EARL LUMLEY *St. Lawrence-Lewis BOCES*
Area 7	GORDON S. PURRINGTON *Guilderland*
Area 8	JOSEPHINE WATTS *Hamilton-Fulton-Montgomery BOCES*
Area 9	KAY MICHELFELD *Warwick Valley*
Area 10	GEORGINE J. HYDE *East Ramapo*
Area 11	IRIS WOLFSON *Nassau BOCES*
Area 12	PAMELA BETHEIL *Longwood*
City of New York	JAMES F. REGAN *New York City*
Large City School Boards	CATHERINE SPOTO *Rochester*
National School Boards Association	ARLENE R. PENFIELD *Clinton-Essex-Warren-Washington BOCES* *Non-voting*

Staff

Executive Director	LOUIS GRUMET
Deputy Executive Director	JAMES V. VETRO
Administrator for Research and Development	JEFFREY M. BOWEN
Research Associate	ROSEANNE FOGARTY
Research Assistant	BARBARA L. BRADLEY
Research Assistant	ANDREA HYARY
Research Assistant	GLEN MCBRIDE
Production Assistant	DIANE REMMERS
Editors	RICHARD L. ORNAUER RITA C. STEVENS SISA MOYO

Gauging Student Performance

A Position Paper of the

NEW YORK STATE SCHOOL BOARDS ASSOCIATION

©1989 by the New York State School Boards
Association, Inc.
119 Washington Avenue, Albany, N.Y. 12210

All rights reserved. No part of this paper may be reproduced in any form or by any means without permission in writing from the publisher.

Printed in the U.S.A.

3M CLG/288

Table of Contents

Summary of Recommendations v

Introduction 1

What are the Purposes of Educational Testing and Assessment? 1
Why Should Testing and Assessment Issues Concern School Boards? 3

Testing and Assessment Policy and the Philosophy Behind It 5

What are the Current Issues Surrounding the Use of Standardized Tests? 7

The Impact on Instruction of Overreliance on Standardized Testing 7
The Psychological and Social Implications of Overreliance on Standardized Testing 10
Misinterpretation of Test Results 13
Problems with Norming 13
Other Inappropriate Comparisons 14

Improving Tests and Testing 16
Testing Higher Order Skills 16
Testing Students with Handicapping Conditions 17
Test Bias 19
Test Administration 20

Testing and Assessment at the State and National Levels 22
New York's State Testing and Assessment System 23
Testing and Assessment Systems in Comparable States . . . 24
National Assessment Systems 25
Testing Results and Funding 26
Recommendations on State and National Testing and Assessment Systems 27

Appendix A 29

Footnotes 31

Bibliography 33

Summary of Recommendations

1. School boards should adopt written policies on testing and assessment, articulating the purpose of tests and other data collection in meeting educational goals and administrative objectives. The primary goal of a testing policy should be to improve individual student learning. A secondary goal should be to assist the state government with the fulfillment of its testing agenda.
2. In developing testing policy, school boards should solicit input from all interested parties, since test scores and other data affect placement, funding, and program decisions.
3. School boards should understand the programmatic and instructional implications of testing in order to interpret test results to the public, and to articulate concerns about testing to state policymakers and test creators.
4. School boards should use their knowledge about testing to avoid the misuse or overemphasis of test results, and the use of inappropriate tests.
5. School boards should be certain that the desire to improve test scores does not become a detrimental influence on instruction, curricula, textbook selection, and the testing program itself.
6. In determining testing policy and the use of test results, school boards must be sensitive to the negative psychological effects of failure to do well on standardized tests. Students' self-concept, motivation, and educational success all are affected by a board's decision about testing.
7. Standardized testing of very young children should not be used as the sole basis for decisions about readiness and placement. The academic mold is too restrictive for children of this age; it should not be used to make children feel deficient or to lower their expectations of themselves.
8. Results of readiness tests should not be used to establish segregated kindergarten classes or keep a child from entering kindergarten.
9. State governments should focus on increasing access to preschool programs, and not on instituting readiness testing programs, which discriminate against children who have had no preschool experience.
10. Districts must be aware of the effect of outdated norms on the interpretation of norm-referenced test results. Every effort should be made to use tests based on updated norms. In addition, publicized results of these tests should include a caveat about the effect of changing norms on accurate interpretation of the results.
11. School boards must understand the flaws inherent in many test score-based comparisons made in today's policy-making arena. The most appropriate and educationally beneficial comparisons are those

that compare an individual student's performance with his/her own performance at a later date. It follows that school districts should be compared to themselves in terms of progress over time.

12. Well-written norm-referenced and criterion-referenced tests have their appropriate places in educational assessment; policymakers must decide which is appropriate for a specific situation.
13. State and local school boards must set goals emphasizing higher order skills and must invest in the development of assessment practices that adequately measure progress toward fulfillment of those goals.
14. In order to provide an appropriate education to all students, school boards must be aware of the alternative testing techniques available to students with handicapping conditions. These techniques enable students with handicapping conditions to participate in testing programs on an equal basis with other students, and to achieve the goals set for them in their individualized education program (IEP).
15. Testing programs must be carefully scrutinized for cultural, gender, and language bias.
16. Alternative systems for distributing scholarships fairly and equitably must be considered and tested.
17. Problems with New York's testing system, including overscheduling of students into too many tests per day, student cheating, and teacher leniency or inconsistency in grading, must be examined and corrected.
18. As the state reviews its accountability system for schools, individual student learning should be a prime consideration. The fairness of evaluative measures, usefulness of data generated, and districts' public image also should be considered.
19. Funding for education should not be linked to test results, but the concept of giving higher performing and improving districts more autonomy in the use of funds should be explored.
20. If a national assessment of education that compares states is to be done, improvements must be made in the assessment tools used and the way comparisons are drawn. An overly simple approach to national assessment will not contribute useful information and could result in misinterpretations and misunderstanding.
21. State and national policymakers should consult local school leaders, who are responsible for implementing educational change, before establishing any broad testing and assessment system.
22. Such systems should use multiple indicators and should concentrate on trends and changes in performance, not on comparing schools, districts, or states.
23. The press and the public must be educated to an understanding of test results and other assessment data.

Introduction

Educational reforms of the 1980s have been accompanied by an increasing emphasis on accountability at all levels. Concern over America's economic and technological future has generated a need to have a better idea of what—and how much—students are learning. Large increases in state aid for education have resulted in a desire to know what the added dollars have purchased. To a great extent, more accountability has come to mean more testing and greater reliance on test results and other quantifiable data.

Testing and assessment have become extremely important components of the educational program. As such, the relevant issues should be of concern to anyone involved in education. The purpose of this position paper is to provide factual and policy-related information that may be of use to school boards as they design or improve their districts' testing and assessment policies, and of use to state officials as they consider new ways to measure and account for changes in district, school, and student performance.

What Are the Purposes of Educational Testing and Assessment?

First, the distinction between testing and assessment should be made and emphasized. Assessment is the broad term that will be used to attempt to determine what educational progress has been made and what improvement is needed by students, schools, districts, states, and nations. Under assessment, testing appears as but one method of assessing progress and problem areas. Although there are many new and innovative ways of measuring progress, testing remains the predominant method; thus, the issues surrounding testing will be the focus of this paper.

There are both *formative* and *summative* purposes of assessment. Formative assessment focuses on the educational experience of the individual student: What are the student's strengths and weaknesses, and what can be done to facilitate improvement? Formative assessment is diagnostic and remedial in nature. Summative assessment constitutes more of a statement of the progress—or lack of progress—toward educational objectives. Summative assessment often serves *administrative* purposes.

A specific assessment program need not be exclusively formative or summative. The same program may serve multiple purposes. The challenge is to develop assessment tools that serve the needs of individual

"A well-designed test, tailored to the curricula, can be an excellent tool for showing a teacher a student's weaknesses and strengths, what remediation—if any—is needed and the extent to which specific material has been mastered."

---

students *and* the needs of districts, states and other administrative entities, without spending more time and money on assessment than on learning.

Testing and assessment are conducted at many levels, ranging from the assessment of progress toward the goals set forth in the individual education program of a student with handicapping conditions, to the National Assessment of Educational Progress, which profiles the condition of education nationwide, and is used to compare the achievement of students in different geographic regions.

Perhaps the most important level at which testing occurs is in the classroom. A well-designed test, tailored to the curricula, can be an excellent tool for showing a teacher a student's weaknesses and strengths, what remediation—if any—is needed and the extent to which specific material has been mastered. This type of diagnostic testing, used extensively in classrooms, is an essential part of the instructional program.

In addition, because of the personal knowledge teachers can (and should) have of their students, teachers can view test scores in their proper context, in addition to students' performance on homework, and their class participation. This is actually the ideal situation: Students are assessed according to a variety of criteria, which reflect different types of learning, thus giving every student a chance to show his/her strengths.

Testing and assessment systems can be used to monitor school building and school district performance. This aggregated data gives administrators an idea of how the school is doing overall, and whether there are problems in particular grade levels or schools.

Data also can be used to compare districts, or schools within a district, although the validity of such comparisons is questionable, as will be discussed later on. Here the assumption often is made that school administrators and teachers can be motivated to work harder to improve

when they know that taxpayers and parents can easily compare districts or schools with one another on measurable performance.

State-mandated testing systems can ensure that districts are complying with requirements to meet certain educational standards. For example, New York State's Regents competency testing program is intended to assure that all students receiving a high school diploma have achieved a minimum standard of proficiency in the basic skills. If students do not score at or above a state reference point on the state's pupil evaluation program (PEP) tests, they must receive remedial instruction.

Traditionally, New York has maintained a strong testing system, spearheaded by Regents examinations. To obtain a Regents diploma, a student must pass these achievement tests based on State Education Department-recommended courses of study in grades 9 to 12. This system, now covering 16 subjects in areas of English, foreign languages, mathematics, sciences, and social studies, historically has distinguished New York from other states.

Other important uses of test data include the selection of college students from the applicant pool and the awarding and allocating of state funds to students, schools, and school districts. For example, New York State Regents Scholarships are awarded to students based on test scores. Another example is the additional state aid provided to school districts based on the percentage of students scoring below the state reference point on PEP tests.

Ideally, no assessment program should confine itself to test scores as the only important measures. Student attendance, disciplinary incidents, dropout rates, enrollment in challenging courses, quantity and quality of extracurricular participation, portfolios containing samples of students' writing and other work, all potentially are useful indicators of student motivation and trends in the school program. Student demographic data also can help fill out the picture of student needs and possible program directions.

Why Should Testing and Assessment Issues Concern School Boards?

An informed school board is the logical bridge between a school district's goals, and the needs and objectives of other stakeholders in education.

As the link between the district and state authorities, school boards satisfy state objectives by implementing state testing programs. In doing so, however, they must understand the programmatic and instructional

implications for students, teachers, and schools. The board must understand and be able to interpret data to the public: parents, taxpayers, and the media. Moreover, school boards also should be able to articulate their needs regarding tests and other data collection to test creators and legislators.

Most important, the school board must know enough about testing to provide, through its written, adopted, and clearly understandable testing policy, a strong framework within which administrators, teachers, and students can carry out the business of learning. Appendix A provides a sample policy.

Despite the fact that the state testing program is extensive, and that before long there may even be a nationwide testing system, school districts should not rely on state and federal authorities to determine their testing policy and program. The goals of these testing programs do not necessarily coincide with the goals of a school board for its staff and students. School boards must have a solid understanding of testing in order to ensure that testing occurring in the schools furthers district as well as governmental goals.

Testing companies, state governments, special interest groups, and the media have been leaders in bringing testing issues to the public and establishing priorities for testing and assessment. Informed school boards with strong, self-initiated student assessment policies should reassert leadership in this area.

Testing and Assessment Policy and the Philosophy Behind It

School boards should have a written testing policy (see Appendix A). A written policy communicates that student testing is an educational issue included in the board's policy-making domain. Asserting the role of the school board in testing policy is especially important because it is an area in which the state already exerts much control. A written policy also communicates the importance which testing is assigned by the board.

A testing policy should set forth the board's basic beliefs about the purpose of testing, thus setting a tone, and providing a framework for more specific policy decisions. Establishing a philosophy of testing and assessment largely will be a matter of balancing the furthering of educational/instructional goals and the satisfaction of administrative/bureaucratic objectives. It also should establish a balance between the use of test scores and the use of other kinds of data for making program decisions.

A testing and assessment policy should have as its primary goal the promotion of individual students' learning. Various strategies can be developed to meet this goal. For example, standardized test results can be

used to identify weaknesses and improve instruction if those results are not filed away and ignored. Efforts can be made to acquire new tests that assess thinking skills in addition to basic skills. District testing systems emphasizing teacher-written tests can be developed. Student measures other than test scores can be given due attention. Education-oriented testing and assessment policy does not eliminate the administrative and comparative uses of data; it merely de-emphasizes the importance of those uses.

While the primary goal of a testing policy may be educational, the state testing programs and their objectives cannot be ignored. It is the board's responsibility to ensure that all testing and reporting requirements are met. Therefore, district testing policy should have as an objective assisting the government with the fulfillment of its testing agenda. This agenda focuses on accountability, funding decisions, and the comparison of results at the building, district, and state levels.

Also to be incorporated into a testing policy are objectives relating to the role of test scores and other data as important district public relations tools. Strategies for reporting and interpreting results to the public, especially to parents, should be laid out.

Finally, students' test results can be important to parents, students, teachers, schools, and the district as a whole. Placement, funding, and program decisions with long-term implications often are made based on these results. Therefore, any policy development certainly should include input from all interested parties.

Again, the most important characteristic of a testing and assessment policy is that its content reflects the board's philosophy about test scores and other data as tools to further the educational goals of all students.

What are the Current Issues Surrounding the Use of Standardized Tests?

The Impact on Instruction of Overreliance on Standardized Testing

In historical perspective, standardized tests represent a significant improvement in the way students are evaluated. For example, before such tests were used, decisions such as college admission were based on subjective criteria. As such, standardized testing has made a great contribution toward achieving equal educational opportunity in this country.

In recent years, however, demand for evidence of improvement has increased greatly the "stakes" associated with test scores. Test scores increasingly are seen by students, teachers, parents, administrators and others as being used to make important decisions that directly affect them. This view results in the growing influence of tests on what is taught and how it is taught.

It should be pointed out that it is the exaggerated importance of standardized test results that negatively affects instruction. Proper use of the tests and results can enhance the educational process.

The main problem with exaggerating the importance of test results is the tendency for classroom practice to react by overemphasizing testable knowledge and its objective evaluation (i.e., multiple choice and fill-in-the-blank tests). While tests of other than factual recall and basic skill areas have been developed, most schools have not yet begun to use them.

Most standardized tests, including those based on state curricula, measure a narrow spectrum of skills. In addition, the type of quick responses required by most tests do not evaluate creativity, the use of important thinking skills, or the ability to express oneself in writing. In short, standardized tests are not good proxies for skill development in many areas, such as problem solving and critical thinking, considered by most to be important components of education.

The overemphasis on testable knowledge and rote expression of that knowledge is manifested in different aspects of the educational program. "Teaching to the test" can mean covering exclusively that material that both teachers and students know is on the test, and even teaching it in the format in which it appears on the test (i.e., multiple choice, fill-ins, etc.). One point, George Madaus reported on testimony delivered at a 1981

National Institute of Education hearing on minimum competency testing:

> . . . (The) principal of a public school in Manhattan said that reading instruction in New York City closely resembles practice in taking tests in reading. In typical reading classes, students read commercially-prepared materials made up of dozens of short paragraphs about which they then answer questions. The materials they use are designed to look exactly like the tests they will take in the spring. (She) further noted that, when synonyms and antonyms were dropped from a test on word comprehension, teachers promptly dropped the commercially-prepared materials that stressed them.[1]

Also affected by state curricula and accompanying tests is the quality of textbooks. Diverse state curricular guidelines and the large, front-end

investments required to develop new texts create pressure to publish texts with broad appeal, if a profit is to be made.

Teachers can be demoralized by pressure to emphasize testable knowledge presented in test format. While a goal of reform has been to enhance the teaching profession, administering workbook activity, fill-in-the-blank exercises, and drilling for tests take up increasing amounts of teaching time. According to education historian Diane Ravitch, the shift in the classroom from teacher control to materials control contributes to the "deskilling" of teachers, converting them from professionals to civil servants.[2]

A final impact of overreliance on standardized testing is the amount of time, better spent on instruction or other activities, taken up by the actual taking of tests and by the weeks or months of drilling before the tests.

The irony of overreliance on test results and the accompanying narrowing of curricula and instruction, is that *test scores will invariably rise!* If the test is changed significantly, scores will decline, and then rise again as teachers and students become familiar with the content of the new test.[3]

Policymakers may point to rising test scores as evidence that the quality of education is improving, and that more students are prepared to function in today's society. However, students can become experts at passing tests, yet may not be able to read, write, or do basic arithmetic. And even if they can read, write and do arithmetic, if their instruction has been test-focused, then much that is important (but difficult to measure) has been ignored. A 19th century, British school inspector insightfully commented:

> Whenever the outward standard of reality (examination results) has established itself at the expense of the inward, the ease with which worth (or what passes for such) can be measured is ever tending to become in itself the chief, if not sole, measure of worth. And in proportion as we tend to value the results of education for their measurableness, so we tend to undervalue and at last to ignore those results which are too intrinsically valuable to be measured.[4]

Many tests students take measure achievement in the basic skills—reading, mathematics, writing. If a primary educational goal is for students to master these skills, and if the tests are valid[5] measures of mastery, then the quality of education may be improved by closely aligning curricula and instruction with the tests which will measure those skills.

For example, the degrees of reading power test that is part of New York State's pupil evaluation program is a highly-sophisticated, nationally-recognized test. If a student's score on this test improves, then in all likelihood the student's ability to read has improved. Thus, if teachers' efforts lead to improved scores on this test, it could be said that teaching to the test has been beneficial.

Measurement-driven instruction, which bases curriculum content on test content, has certain advantages that include:

- making the goals of instruction explicit
- focusing teacher and student efforts on well-defined targets
- making standards clear and uniform, thereby making accountability at all levels clearer and more objective
- giving the public concrete information on how well the schools are doing[6]

Tests that are viewed as important enough to drive curriculum also have the power to initiate innovations in curriculum. For example, Madaus pointed out that the New York State Education Department was not successful in changing the emphases of the foreign language curricula until the corresponding changes had been made in the foreign language Regents examination.[7]

Despite some advantages, measurement-driven instruction has limited positive educational effect. In fact, the likelihood of negative ramifications seems great in all but a few learning situations. Measurement-driven instruction is merely one way to clarify goals and develop uniform standards, efforts which should characterize any sound assessment program.

The Psychological and Social Implications of Overreliance on Standardized Testing

Testing and test results have assumed a position of great importance in education today. The greater the rewards and sanctions associated with success and failure, the greater the impact will be on those involved. In this context of overemphasis on test results, it is easy to understand how students' self-concept, motivation, and educational plans could be affected negatively by consistent failure to perform well on standardized tests. The psychological effects of failure can increase the likelihood of dropping out.

Considering the importance students attach to test results, Andrew Strenio, author of *The Testing Trap*, asked: "How many (children) can shrug off this judgment from on high as just an estimate subject to error of some of their skills at a particular moment? Can we really expect these children to display the sophistication and skepticism about test results so lacking in many adults who use standardized tests?"

Placement in remedial classes or programs is often the consequence of poor test results. Misplacement may occur sometimes. The stigmatization

and negative aspects of such placement are well-known. Children in "slow" classes are seldom expected to perform well. The least experienced and least competent teachers often are assigned to these classes. Once a child is in a remedial track, his or her placement may be viewed as permanent. Frustration with this situation can lead to withdrawal, absenteeism, and dropping out. One must be cognizant of the long-term effect on student retention of relying too heavily on standardized test results.

The goals of a student's instructional program should be clear. The program should challenge and interest each student, yet expectations should not be unrealistic. The assessment of student progress should be valid and reliable, and students should know what is expected of them. Tests appropriate for diagnostic purposes can—and should—be used, along with other measures, to identify students having trouble as well as areas of weakness in curricula and instruction.

Spotting problem areas and providing appropriate assistance before students get frustrated and turned off helps maintain motivation, and can help prevent dropping out.

Under these conditions, testing can enhance self-esteem in students by mapping progress in a concrete way, and providing much opportunity for success. However, just as a testing program can motivate and promote self-esteem, its misunderstanding and misuse can cause irreparable damage to student academic motivation and psychological well-being.

Especially crucial are the psychological and educational implications of testing very young children. As part of their reform efforts, several states have instituted mandatory readiness testing for children entering kindergarten. New York State, however, does not have a readiness testing program.

Children who are tested to enter kindergarten are being tested too early and failed too soon for no justifiable purpose. A complete diagnostic evaluation must be conducted in order to evaluate a child's strengths and weaknesses.

In many cases, the Gesell test is used to the exclusion of any other instruments. This test takes 15 minutes and can be administered by teachers who have participated in a five-day course. In other words, teachers with no background in special education are making decisions that will likely affect childrens' entire academic career. The validity and reliability of the Gesell test have been questioned, as has the appropriateness of all readiness testing.

In some districts, children deemed not ready for kindergarten are placed in a remedial kindergarten class. Class placement based on the assumption that remediation is needed at such an early age contradicts widely accepted learning theory that young children develop at different rates. The academic mold, especially in early childhood programs, is too restrictive; it should not be used to make children feel inferior or deficient.

In addition to readiness testing, the use of standardized tests to evaluate the performance of very young children (kindergarten through third grade)

"Results of readiness tests should not be used to establish segregated kindergarten classes or keep a child from entering kindergarten. Many of the behaviors that constitute 'readiness' are taught in today's preschool programs that are unavailable or inaccessible to many children."

---

has become more widespread. Testing at such an early age is often counterproductive; young children are unschooled and erratic test-takers; testing them results in too much mislabeling.

When Mississippi decided to halt the statewide, standardized testing of kindergarten students, an evaluator of the program said teachers were letting the test become a curriculum guide, contrary to the program's intentions. Evidently, teachers felt their students' performance reflected their own ability; they began to use formal pencil and paper drills to prepare children for the test. This practice was not consistent with the leading aim of the kindergarten program—to help students develop a positive self-concept. The program's guidelines stated children should not be pressured to perform in ways that were not developmentally age-appropriate.

The National Association for the Education of Young Children and the National Association of Early Childhood Specialists in State Departments of Education have issued statements critical of standardized testing of young children. They claim there is strong evidence that kindergarten retention, the fate of some low scoring children, is educationally and psychologically harmful and does not appear to improve chances of school success.

Results of readiness tests should not be used to establish segregated kindergarten classes or keep a child from entering kindergarten. Many of the behaviors that constitute "readiness" are taught in today's preschool programs that are unavailable or inaccessible to many children. Readiness testing discriminates against children who have not been in preschool. Instead of readiness testing, states should focus on increasing access to preschool programs.

Used flexibly, and with other diagnostic tools, early standardized testing can spot and address student learning problems. Once again, it is not the use of tests that causes damage, but the overreliance upon the results in making educational decisions.

Misinterpretation of Test Results

Problems with Norming

The essence of a norm-referenced test is that it compares an individual's performance to that of a national sample of students, the norm group. The testing company administers the test to a national representative sample of students before the test is distributed to districts. Thereafter, for as long as that edition of the test is in use, all scores are compared to the norm group's scores. The score earned by the middle scoring pupil in the norm group—the 51st of 101 children, for example—becomes the grade level score. Individuals who correctly answer more questions than the middle pupil are said to score above grade level, while those who correctly answer fewer questions than the middle pupil are said to score below grade level.

Other types of scores can be derived from norm-referenced tests, but they all convey an examinee's relative standing in a defined group of other examinees. Since the tests are designed to distribute results in a specific way, it is totally unrealistic to expect everyone to score at grade level on norm-referenced tests.

Correct interpretations of norm-referenced test scores, and of statements made based on those scores, require an understanding of the problems associated with norming. First, the comparison of any group of scores to those of the norm group is meaningful only to the extent that the two groups are similar. In other words, if a group of students is not like the norm group in terms of socioeconomic variables, achievement range, test-wiseness, etc., the norm group will not provide a meaningful comparison.

Public attention focused on the problems with norming when a group headed by West Virginia physician John Cannell published the results of its inquiry into nationally-normed, standardized testing.[8] Dr. Cannell stated that all of the states and most of the districts reported test scores ranked above the national average! The booklet was nicknamed the "Lake Wobegon" report after Garrison Keillor's mythical community in which "the women are strong, the men are good-looking, and all the children are above average." According to critics, Cannell's report contained errors; however, notes social scientist Daniel Koretz:

> In my opinion, there can be no doubt that current norm-reference tests overstate achievement levels in many schools, districts, and states, often by a large margin. To my knowledge, none of Dr. Cannell's critics have disagreed with this judgment.[9]

One of the causes of the "Lake Wobegon effect" is the use of outdated norms. If an edition of a test is used for a period of seven years, and during those seven years achievement is increasing on the average nationwide, then the national norm to which scores *should* be compared is actually higher than the original norm established.

The obvious solution is to update norms every year or two years. In addition, published results of these tests should include a caveat about the effect of outdated norms on accurate interpretation of the results.

Other Inappropriate Comparisons

Other explanations of the exaggeration of achievement are not directly related to norming but involve other inappropriate comparisons. For example, districts often choose those nationally-standardized tests which most closely parallel district curricula. At the extreme are districts that hire the testing company to customize the test, thus tailoring it even more closely to the curricula. Obviously, it is inappropriate to compare performance on a customized test to that of the norm group that did not take a customized test. Similarly, if "teaching to the test" is taking place, then performance of those coached should not be compared to the norm group, which did not receive coaching.

Unfortunately, misunderstanding is widespread among policymakers, administrators, media, parents, and others. The comparisons that these "users" are demanding be made are doing serious damage to the quality of education in this country by corrupting what goes on in the classroom, and by directing attention away from the truly important issues.

Many of the comparisons being made are invalid because they compare apples and oranges. It is inappropriate to compare the performance of any two units (students, schools, districts, states, countries) that are inherently different before assessment begins.

It is noteworthy that many districts work hard to break the link between socioeconomic background and academic achievement. Their efforts are based on the essential belief that every child can learn. Nevertheless, it is unfair to compare the test scores of students, schools, and districts into which vastly different levels of resources have been invested.

Another apple to orange comparison is one that appears on former United States Education Secretary William Bennett's wall chart.[10] On the wall chart, the average Scholastic Aptitude Test (SAT) scores of all the states are compared. The result is that states in which a small percent of the students take the test have high average scores, while states like New York, in which nearly 72 percent of the students take the test, had relatively low average scores. In the former states, only the "cream of the crop" take the test; in New York, the test-taking population is much more diverse.

Experts continue to build statistical bridges to account for the population differences and allow for better comparisons. At the same time, politicians in states with the highest average scores—for example, in Iowa where only five percent of high school graduates were tested— like to attribute the scores to the high-quality education received by students in that state.

Comparing schools, districts, states, and countries that have vastly different educational goals and standards is yet another instance of comparing apples and oranges. For instance, it would be unfair to compare

the mathematics scores of a district whose primary educational goal was to prepare its students for the 20th century by concentrating heavily on mathematics skills, to the mathematics scores of a district whose primary goal was to implement a higher order skills program that subordinated subject skills improvement to class discussion and analytical thinking.

Within reason, educational goals and criteria should be—and are—set by local school boards according to the needs and values of the community. The inevitable variety resulting from this form of educational governance invalidates many of the aggregate comparisons that are so popular in today's policy-making arena.

The types of comparisons discussed above are flawed; they also, for the most part, do not help students learn better. The most appropriate and educationally-beneficial comparisons are those that compare an individual student's performance with his or her own performance at a later date. It follows that schools and districts should be compared to themselves in terms of progress over time toward the fulfillment of educational goals.

Improving Tests and Testing

Testing Higher Order Skills

The current reform effort must expand to include the teaching and testing of higher order skills if it is to help the educational system adapt to a changing society. A report of the National Governor's Association concurred:

> Such skills include the ability to communicate complex ideas, to analyze and solve complex problems, to identify order and find direction in an ambiguous and uncertain environment, and to think and reason abstractly. Because workers in the future will experience rapid changes in both work technologies and jobs themselves, students will also need to develop the capacity to learn new skills and tasks rapidly. This will require a deep understanding of the subject matter students study, and an ability to apply this knowledge in creative and imaginative ways, in novel contexts, and in collaboration with others.[11]

There is nationwide evidence that students perform more poorly on tests of higher order thinking skills than they do on tests of the basics. For example, extremely few students could offer even rudimentary support for their opinions on a National Assessment exam, part of which asked students to interpret or evaluate literary sections.[12] Although the rapidly changing society and workworld are requiring the acquisition of more highly developed skills, those skills actually are being de-emphasized in the classroom. The effort to quantify educational productivity through the use of standardized tests, usually of basic skills, is discouraging school systems from truly adapting to the demands of a changing society.

Contrary to what many think, thinking skills *can* be taught and evaluated. They can be taught through programs separate from other curricular areas, and/or they can be infused into every area of the curricula. Examples of the latter include teaching problem identification, goal setting, self-monitoring of progress, and self-evaluation; belief in thinking, experimentation, and persistence; use of a "thinking journal" to help students become more aware of their thought processes;[13] use of "wait time" after a question has been posed; requiring students to defend their reasoning against different points of view, describe how they arrived at answers, develop their own questions, summarize other students' responses.[14]

An example of the former is the Talents Unlimited Program, in operation in several schools across the country. At one elementary school, the program "focuses on critical and creative thinking, invites children to become active learners rather than passive receivers, and enables teachers to function as facilitators rather than disseminators of information." One

Talents Unlimited activity includes the planning and execution of the marriage of Mr. Q and Ms. U, complete with tuxedo, veil, and stereo![15]

Thinking skills programs can emphasize analytic problem-solving, creative thinking, or critical thinking; every program can have its own definition for each of those terms. District educational goals should determine the thinking skills program that is developed or chosen.

The assessment of higher order skills is still in its infancy. Although some thinking skills assessment instruments already exist, most consist of multiple-choice or short answer questions. According to some experts, the "one correct answer" approach to teaching and learning is not compatible with learning to develop one's thought processes.

Thus, while multiple-choice tests are easy and relatively inexpensive to implement, they are not adequate measures of higher order skills. The National Governor's Association's recommendations for higher order skills assessment are as follows:

> . . . assessment tools are needed which require students to synthesize, integrate and apply knowledge and data to complex problems. They should present tasks for which no one answer is right, but for which a range of solutions may be in order. These may take the form of essays, projects, or other demonstrations of competence . . . . Scoring systems which rely on expert judgment rather than simply checking for the correct answer should be part of these assessment tools.[16]

What students learn is at least partly a function of what schools are expected to teach them. *State and local school boards must set goals emphasizing higher order skills and must invest in the development of assessment practices that adequately measure progress toward fulfillment of those goals.* The commitment to helping students learn to think may be a costly one, involving program investigation and/or development, development of assessment tools, staff training, and the classroom time needed for the discussion and reflection that foster better thinking.

The price of investment in the teaching and testing of higher order skills will not be as high as the long-term costs to society and economic productivity of not making that investment.

Testing Students with Handicapping Conditions

Alternative testing techniques should be, and in New York State are, made available to students with handicapping conditions. The majority of students with handicapping conditions have physical, emotional, or learning disabilities which, if addressed through appropriate special educational programs, should not preclude them from meeting the requirements for a local or Regents diploma. Alternative testing techniques enable students with handicapping conditions to participate in testing programs on an equal basis with other students and to achieve the goals set for them in their IEP.[17]

"School board members should be familiar with available alternative testing techniques and the way in which they are matched appropriately with students. This familiarity is essential if the board is to fulfill its main responsibilities regarding alternative testing."

New York State is relatively advanced in this area. In New York, several sections of the commissioner's regulations provide for the availability of alternative testing to students with handicapping conditions. The regulations state that "each student with a handicapping condition . . . shall have access to the full range of programs and services, to the extent that [they] are appropriate to such student's special educational needs."[18] And, "instructional techniques and materials used by schools shall be modified to the extent appropriate to provide the opportunity for students with handicapping conditions to meet diploma requirements."[19] Regulation also provides that students identified by a committee on special education (CSE) as having a handicapping condition may be provided with alternative testing procedures,[20] and that the CSE should include a listing of testing modifications in the students' Phase I IEP.[21]

Examples of alternative tests and testing include Braille tests for the blind, additional completion time for students with impaired manual skills or learning disabilities that require additional processing time, and receiving test directions using auditory amplification devices for the hearing-impaired.

The CSE considers the needs of each student in making recommendations to parents regarding the use of alternative testing procedures. Since virtually all students could benefit from alternative testing techniques (i.e., additional testing time), the emphasis must be on determining necessity for modification rather than potential benefit.

School board members should be familiar with available alternative testing techniques and the way in which they are matched appropriately with students. This familiarity is essential if the board is to fulfill its main responsibilities regarding alternative testing. One of these is to establish district policies in this area and to ensure that the policies and relevant procedures are well understood by district personnel, parents, and CSE members. The other responsibility is to review the recommendations made by the CSE.

Others directly involved with students with handicapping conditions are the committee on special education, school principal (responsible for implementing the district's policies in this area), special education teacher, regular teacher, and parents. Ensuring that each child with a handicapping condition(s) receives an appropriate education, including an appropriate testing program, requires the cooperation and efforts of all these groups.

Test Bias

The inappropriate use of tests can create bias against certain populations of students. For example, a mathematics test given to a student with limited proficiency in English may be biased against the student; her/his score on the test may not reflect knowledge of mathematics but rather an inability to understand the testgivers' directions.

One serious consequence of applying standards for English-speaking students to non-English-speaking students is misclassification. A study of learning disability placements in Colorado showed that among the students so labeled, only 43 percent demonstrated actual learning disabilities. The other 57 percent included slow learners, emotionally-disturbed students, and non-native English speakers.[22] Inconsistent labeling of students and disagreement about the definitions of labels contribute to this problem. It may be alleviated partially by the availability of alternative testing procedures (i.e., native language tests).

In New York State, the Bureau of Bilingual Education of the State Education Department is involved in an ongoing effort to make the testing of limited-English proficient students fair and equitable as possible.

The Regents competency mathematics and native language writing tests are available in 29 languages. These writing tests are not merely translated from the English. They were developed by specialists in the language and culture who, in addition to translating, attempted to make the tests less culture-bound. For example, students who may have had terrible and violent experiences before leaving their country would not be asked to describe "life in their homeland" on the writing test. Similarly, tests for immigrant, teenaged students from Moslem countries like Afghanistan would not refer to dating behavior.

Regents competency science and social studies tests are available in several languages as well. The third- and sixth-grade PEP mathematics tests are available in four languages. In addition, limited-English proficient students who have not received two full years of English language instruction at the time of testing may be exempted from all state tests unless one test is available in their native language.[23]

The dramatic increase in the importance attributed by schools and districts to test scores has had negative effects on many aspects of testing and assessment. One negative effect is the temptation to limit the participation of students whose anticipated low scores will lower the average scores. In light of this, the availability of native language tests and

the option for exemption should be viewed as double-edged swords. In the course of affording equal testing opportunities, one must not lower expectations and the quality of education for limited-English proficient students.

Bias inherent in the content of tests should be eliminated entirely. For example, test questions based on scenarios or language familiar only to middle- and upper-income examinees may affect the scores of lower-income students, regardless of their knowledge of the subject being tested. Examples of such bias include references to cricket or sculling.

An issue that has attracted public attention is an alleged sex bias in testing. For the past several years, the SAT was used to rank students for the purpose of awarding Regents Scholarships, traditionally awarded for academic achievement and excellence. However, use of the SAT resulted in a disproportionate number of scholarships being awarded to male students. The assumed reason is that females tend to take fewer math and science courses than males in high school.[24]

Reacting to critics, the state Legislature changed the basis upon which the Regents Scholarships were awarded to include classroom grades as well as SAT scores. This resulted in an increased number of female students receiving scholarships, since females generally receive higher classroom grades than males. The use of grades is an unacceptable basis for scholarship distribution. Standards for grading vary greatly across districts. At issue is how to award achievement-based scholarships most fairly and equitably.

Changing the tests for everyone would lower standards for everyone and would constitute a "Band-Aid solution" to the deeper problem of bias in the society. Girls often are steered away from math- and science-oriented activities throughout childhood, both at home and at school. As a result, they take fewer math and science courses, and are not as strongly encouraged to perform well in those areas as boys; but changing the tests clearly is not the solution. Nor should what is essentially a quota system have been established for awarding scholarships, in the effort to thwart the effect of the bias.

As one editorial writer noted, it is true that the legislated change resulted in more females winning Regents scholarships; it is also questionable whether the students now winning the scholarships are really the state's highest achievers.[25]

It has been suggested that a new statewide examination be developed to replace the current system. Any new system should 1) ensure equity of distribution among school districts, 2) maximize local control and policy discretion, and 3) make assessment serve instructional purposes.

Test Administration

Literature on the role of testing in education today often quotes teachers who claim that the amount of time students spend taking standardized tests is excessive and growing.[26] In some cases, exams are scheduled back-

to-back so students have little time in between exams to prepare, not to mention the stress and exhaustion such scheduling can cause.

In New York State, for example, it is possible for a student, depending on the particular Regents examinations being taken, to sit for three exams in one day, each of which is three hours long. One of the difficulties faced by districts in establishing creative district assessment systems, is that any new tests introduced would add to the already-heavy schedule of state and federal tests. Finally, it is important to remember that time spent on testing often includes days, weeks, even months of coaching for the tests, and not just the time spent actually taking tests.

As has been stated throughout this report, there are many serious and unfortunate consequences of the heightened emphasis on school and district average test scores. One of these consequences is the temptation to cheat. When tests are not stored securely before administration, they can be accessed and distributed and the actual test items taught to the students. Obviously, this invalidates the results. Some tests are not changed over a period of several years, so that test contents can be remembered and taught to students even if the tests are kept physically secure before administration.

In New York State, the well-reputed Regents examination program has long had highly-effective security procedures, while PEP tests have not. For example, Regents examinations are delivered to school districts in secure metal boxes, while PEP tests are not. In addition, the procedures for administration of Regents examinations are stricter than for PEP tests.

In the past, the PEP test served mainly as a diagnostic instrument, although a limited amount of state funds has been associated with test performance. However, publication by the State Education Department of a list ranking schools in order of their PEP test results, and the sanctions associated with poor performance, have turned PEP into a "high stakes" testing program. A corresponding need to strengthen the security procedures has developed.

The Board of Regents has approved a plan for enhancing PEP test security, including the purchase of metal shipping boxes (a possible one-time cost of $500,000) and the development and printing of annually-revised tests.

Increasing the stakes associated with school and district performance on standardized tests also has exacerbated problems with the grading of tests. In New York State, while the extent of the problem is not yet clear, it is certain that the problem exists.[27] When what are perceived as negative ramifications are associated with poor performance, the temptation is great to have as many students as possible do well. Not only does cheating in this way invalidate the test results, but more important, it cheats the student whose properly-graded exam would have resulted in receiving needed remediation. As with test security, the grading of the Regents examinations has been more carefully monitored by the State Education Department than has the grading of PEP tests.

Testing and Assessment at the State and National Levels

If, as was previously stated, testing policy has as its primary goal the promotion of individual students' learning, what is the purpose of state and national testing and assessment? The arguments usually presented are that test scores and assessment data are necessary tools of policy-making, that elected officials need them to justify educational expenditures, and that the knowledge they provide can improve educational quality in the entire system.[28]

The basic question that recurs today is whether, in the process of meeting those needs through state and national testing and other data collection, the individual student's learning is promoted or hindered. Just as the individual student or teacher can be affected negatively by overreliance on standardized testing or misinterpretation of test results, so can a school building or school system. An Office of Educational Research and Improvement (OERI) task force cited a series of trade-offs that must be made when a state accountability system is set up:

- state accountability vs. local autonomy
- the need for statewide comparability vs. the need for local ownership
- the need to know how well schools are doing in overall performance (excellent, fair, poor) vs. the need for other information (on writing skills, knowledge of science, minority success, teacher quality) needed to fine-tune diagnosis and plan improvements
- the usefulness, thoroughness, and fairness of collecting and analyzing new data vs. the burden of paperwork, time, and effort it takes to do so
- the economy of easy assessment methods (like nationally-normed standardized tests of basic skills) vs. the importance of measures of other aspects of school performance, like problem-solving skills and participation in the arts
- a state's desire to assess local schools and help them improve vs. consideration of the state budget, staff expertise, political climate, and technology available to do so[29]

Most of the above choices also are faced in establishing a national testing system. The list of trade-offs makes it clear that the design of a state or national testing and assessment system must be sensitive to possible negative impacts. The goal must be to make the system practical, not punitive.

New York's State Testing and Assessment System

The constitutional responsibility for education rests with the states, and, therefore, state attention to educational data collection is appropriate. New York has one of the longest-standing systems for assessing student performance statewide in its 100-year-old Regents testing program. Today, student scores on those tests, along with a bevy of other test scores and data, must be reported annually to the public by each school district in a Comprehensive Assessment Report (CAR). The report includes:

- pupil evaluation program (PEP) test scores for reading and mathematics at grades three and six, and for writing at grade five
- program evaluation test scores for social studies at grade six
- preliminary competency test (PCT) scores for reading and writing at grade eight or nine (local option)
- Regents competency test (RCT) scores for reading, writing, mathematics, and science
- scores on the various Regents comprehensive exams in academic courses

In addition, districts are required to report student enrollment by grade, dropout and attendance rates, data on certificates and diplomas awarded and students transferring into equivalency diploma preparation programs.[30]

For the past several years, districts that were identified as "most in need of assistance" on the basis of the CAR, were required to prepare a Comprehensive School Improvement Plan (CSIP), which had to be implemented within a year and monitored periodically by the State Education Department.

New York's accountability system, as embodied by the CAR and the CSIP, has raised many of the trade-off issues cited by the OERI task force. The State School Boards Association, as well as the CSIP schools, has questioned the fairness of the measures used and the usefulness of the data generated, as well as the negative public image resulting from being placed on a list of schools "in need of assistance." Even schools not on "the list" cite the reporting burden and the dampening effect on local autonomy in setting educational goals. New York State's assessment system is being reviewed. Changes being considered include:

- requiring *all* schools to have a comprehensive long-range planning process
- setting standards of excellence as well as minimum standards, and the "negotiation" of progress indicators by the department and individual schools[31]

Testing and Assessment Systems in Comparable States

Although New York's system for statewide testing and assessment is comprehensive, it is by no means the only type of state system extant. The following examples illustrate different models for state approaches to testing and assessment.

New Jersey uses what might be called the "compliance model." Using 43 different indicators of school performance, the state classifies districts into two levels, level I being satisfactory and level II being in need of improvement. Level II districts must submit and implement corrective plans, and those that fail to reach level I in subsequent monitoring are classified as level III. Level III districts are subject to takeover by the state, as the case of Jersey City illustrates.[32]

California uses a "performance model" which in a somewhat less punitive way attempts to classify and rank schools, and set expectations for performance. A key feature is its use of demographic and socioeconomic data to group schools into strata that establish a "band of expectation" for each school's improvement. In a survey by the Council of Chief State School Officers, less than half the states responding reported that they made use of socioeconomic and demographic data in their assessment system.[33]

Minnesota provides a "diagnostic model," focusing less on compliance with statewide standards and more on local diagnosis of individual student needs. Similar to outcome-based education, the Minnesota approach encourages locally-developed instruments closely linked to local curricular objectives, and a diagnostic/remedial approach to test results. State instruments also are used; each district is required to administer annually, at three grade levels, a state-developed assessment test in at least one subject area. Districts are required to report results to the public, but comparisons between districts are discouraged. Because Minnesota students consistently perform well on national assessments, the state has resisted going to more stringent statewide monitoring systems that might discourage local initiative.[34]

In a fourth example of a statewide innovation, Pennsylvania has attempted to go beyond the monitoring of basic competencies by developing a state assessment of thinking skills. The test, which is administered in grades 5, 8, and 11, is voluntary for school districts. In a given year, between 30 and 45 percent of Pennsylvania's districts elect to use it.[35]

The above examples describe some of the existing variations on statewide accountability systems. Attempts to measure thinking skills, incorporate socioeconomic data, and encourage locally-developed instruments are all steps in the right direction.

Other innovations in assessment have been summarized in a National Association of Secondary School Principals (NASSP) publication. These include the use of "participation indicators" such as participation in

extracurricular activities, self-help or support groups (teen pregnancy, drug abuse, etc.), and attendance and dropout rates. The assessment of writing achievement should receive more attention, as should the disaggregation of test results in order to track progress over time, between schools, and between groups of students.[36] While New York's system is comprehensive and incorporates some of these examples, there may be important factors that have not been considered in its design.

National Assessment Systems

It might be argued that the United States already has a national testing and assessment system —the National Assessment of Education Progress (NAEP). Since 1969, the federally-funded NAEP has tested about 100,000 students nationwide every two years in reading, writing, mathematics, and other subjects. In the past, the laws regulating NAEP did not allow interstate comparisons. Authorization for such comparisons was granted this year, however, and in 1990 a trial, eighth grade, math assessment will be conducted "with the purpose of determining whether such an assessment yields valid, reliable, state representative data."[37] At least 20 states have volunteered to participate in the pilot test. This expansion of NAEP to allow interstate comparisons reflects business and policymakers' desire to know "how the states are doing."

Other forms of national assessment include the aptitude tests administered by the Educational Testing Service, and a national "wall chart,"

published by two successive secretaries of education ostensibly to compare the educational performance of the 50 states. The results of these tests and comparisons, however, generate more questions than they have answered, and thoughtful critics have been asking how valid or useful this type of assessment is.

For example, no attempt has been made to take into account differences in state curricula or methods of instruction, or the reasons behind those differences. In the absence of a national curriculum or agreed on canon of knowledge, such as exists in some smaller countries, these differences cannot and should not be ignored.

No attempt has been made to factor in differences in the total coursework preparation of students taking the various tests. Even more serious, differences in the populations taking the tests from state-to-state have been disregarded in the wall chart. Various researchers have used correction factors, which may or may not themselves be valid, to show how major reversals in state rankings can occur.

For example, Steelman and Powell showed that SAT state rankings in 1984 placed New York in position 35.5. However, 59 percent of New York seniors elected to take the tests, compared to 2 to 4 percent in some other states. When a correction factor was incorporated to take these differences into account, New York's national ranking was 5.5.[38]

These are just a few of the state and regional differences that cast doubt on the validity of existing attempts at a national assessment system.

A report by the Congressional Budget Office in 1987 underlines the weakness of those attempts. The report cites the limitations of standardized tests of basic skills, the complex of educational and noneducational factors that vary across states and regions, and the unavailability of data necessary for valid comparisons. It warns against simple-minded solutions to problems of educational performance that are essentially complex.[39]

Testing Results and Funding

One of the most vexing questions related to state and national educational assessment systems is whether educational aid should be linked to a state's, school district's, or school's performance on a given set of tests or other measures. Both a positive and a negative response have good justifications. There is general recognition that poor performance is often related to deprivation of resources, either on the part of the schools or of the children they serve. Thus, the rationale for compensatory aid, sparsity or density aid for inner city or rural districts, etc., is strong. Further, the questionable validity of test results as a measure of school effectiveness argues against linking aid to performance.

On the other hand, the psychology of incentives suggests that schools should be rewarded for good performance, and that recognizing improvement in a tangible way is the best method for encouraging more improvement. Still, given the limitations of educational funding and the

great needs of many schools and student populations, can government afford to fund a reward system for high-performing schools, especially those with adequate resources already?

Educational aid must be based on a philosophy of leveling, of attempting to correct social and educational inequalities and thereby raise performance. It should not be assumed, however, that when a school with a deprived student population succeeds in raising its student performance, that the need for funds decreases.

The student population of subsequent years will arrive with just the same needs, and the school will have to meet them. Depriving a school of aid when it improves its performance may only discourage improvement and, therefore, penalize students. Thus, performance itself, either good or bad, should not be the factor that drives the amount of funding a school receives.

The concept of having performance determine, not the amount of funding, but the degree of autonomy in the use of funding may be the solution. It would be appropriate to be more prescriptive about a school's use of its funding during a low-performing period, and to loosen restrictions when it becomes clear that the school has charted a successful path to improvement for itself.

One caveat on prescriptive aid for low-performing schools should be added, however. In some cases, legislated aid packages can become so prescriptive that aids are fragmented into amounts too small to be useful. When an individual school or district receives its allotment, it discovers that no piece of the funds received is sufficient to mount any major effort at reform. The result—business as usual, and frustration of the goal of school improvement for which the aid was originally designed.

The bottom line on assessment and funding is that the basic goal of educational funding must continue to be the elimination of inequalities in resources, and that, however tempting the idea of monetary incentives may be, the science of testing and assessment is still too inexact to be the basis for assigning rewards.

Recommendations on State and National Testing and Assessment Systems

The questions raised and the models cited in the preceding sections suggest some generic recommendations that apply equally to New York's state accountability system and to any existing or proposed national system for assessing educational performance.

First, it seems obvious that an assessment system should have a central organizing theory. While such a recommendation seems self-evident, the fact is existing data and measurement instruments frequently drive the design, purpose, and effects of the system, not the reverse.

Policymakers should ask, "What is the system meant to do?" and an important corollary question is, "Who should design it?" If the needs of those who are responsible for implementing change, the local school

leaders, are not consulted, then it is unlikely that the system will be useful to them.

Second, a valid assessment system should use multiple indicators, not merely standardized test scores of basic skills, and those indicators should be derived from agreed on goals for education. If education is charged with doing much more than teaching reading, writing, and math, its success cannot be judged by limited measures of those skills, nor do those measures give any adequate guidance for future directions.

Third, an assessment system must attempt to correct for population differences to sift out the actual effects that the schools are having. If schools are to be judged, it must be on the basis of "value added" by the school, not on the basis of scores that ignore differences in students' starting points. Furthermore, the focus should be on change and trends, rather than on performance at isolated points in time.

Finally, an assessment system must be designed with careful consideration of how its data are to be used. If comparisons are made, they must be valid ones, and both the press and the public must be adequately educated to understand their meaning.

State and national policymakers contemplating testing and assessment systems should ask whether their systems, existing or planned, truly serve the goals of education and the needs of individual students, and whether they effectively promote or block educational improvement.

Appendix A

Sample Policy
4720

Testing Programs

The Board of Education believes that testing and assessment data can provide a meaningful source of information about district curricula and student achievement. Improved student learning is the primary goal of both standardized testing and teacher-designed tests currently in use in the district. In addition, all assessment instruments should help accomplish the following objectives:

1. to evaluate strengths and weaknesses of the current curriculum and methods of instruction;
2. to provide one means to evaluate student growth through individual, interdistrict, and intradistrict comparison;
3. to provide teachers with diagnostic information which will enable them to better address the instructional needs of their students; and
4. to provide a basis for longitudinal study of student achievement.

Information gained through the use of testing programs will be used to design educational opportunities for students to better meet their individual and collective needs. The Board views this purpose to be a primary function of schools.

The Board recognizes that tests provide only a limited source of information, and will, therefore, be used only in conjunction with all other information known about a student or to assist the student in improving his/her work. No test(s) will be used as the sole basis for decisions regarding an individual student's readiness or class placement.

Records of the results of standardized tests shall be maintained in accordance with the Board's policy on student records.

Cross-ref: 1120 School District Records
4721 Test Selection and Adoption

[This policy was developed by the Policy Services Department of the New York State School Boards Association.]

Footnotes

1. George Madaus, "Test Scores as Administrative Mechanisms in Educational Policy," *Phi Delta Kappan*, May 1985, pp. 611-617.
2. Diane Ravitch, "The Uses and Misuses of Tests," *The College Board Review*, Winter 1983-84, pp. 23-26.
3. George Madaus, "Testing and the Curriculum: From Compliant Servant to Dictatorial Master." In Tanner, *Critical Issues in the Curriculum*, NSSE Yearbook, 1988.
4. E. G. A. Holmes, *What Is and What Might Be: A Study of Education in General and Elementary in Particular*. As cited in Madaus, Testing and the Curriculum.
5. Validity here refers to the extent to which a test actually measures what it is intended to measure. For example, does a multiple-choice reading comprehension test measure how well a student has learned to read, or does it measure how well that student has learned to take multiple-choice reading comprehension tests, or how much they know about the particular topics discussed in the passages?
6. George Madaus, "Testing and the Curriculum: From Compliant Servant to Dictatorial Master." In Tanner, *Critical Issues in the Curriculum*, NSSE Yearbook, 1988.
7. Ralph W. Tyler, "The Impact of External Testing Programs," in *The Impact and Improvement of School Testing Programs* ed. W. J. Findlay. As cited in George Madaus, "Testing and the Curriculum: From Compliant Servant to Dictatorial Master." In Tanner, *Critical Issues in the Curriculum*, NSSE Yearbook, 1988.
8. Dr. John Jacob Cannell, *Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States are Above the National Average*, 1987.
9. Daniel Koretz, "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?" *American Educator*, Summer 1988, p. 11.
10. *State Education Statistics: Student Performance and Resource Inputs*. U.S. Department of Education, Office of Planning, Budget, and Evaluation, February 1988.
11. Michael Cohen, *Restructuring the Education System: Agenda for the '90s*, National Governor's Association Center for Policy Research, December 1987.
12. Ibid.
13. John Barell, Rosemarie Liebmann, and Irving Sigel, "Fostering Thoughtful Self-Direction in Students," *Educational Leadership*, April 1988, pp. 14-17.
14. Jay McTighe and Frank Lyman, Jr., "Cueing Thinking in the Classroom: The Promise of Theory-Embedded Tools," *Educational Leadership*, April 1988, pp. 18-24.
15. Edmund Barbieri, "Talents Unlimited: One School's Success Story," *Educational Leadership*, April 1988, p. 35.
16. Ibid.
17. New York State Education Department, Office for Education of Children with Handicapping Conditions, *Alternative Testing Techniques for Students with Handicapping Conditions*, May 1986.
18. Part 100 of the Regulations of the Commissioner of Education, New York State Education Department, Section 100.2, Subsection s1.
19. Ibid., Subsection s2.
20. Ibid., Subsection g.
21. Ibid., Section 200.4, Subsection c2.
22. Educational Testing Service, *Testing, Equality, and Handicapped People*, 1988.
23. Report on "Services to Limited English-Proficient Students 1984-1988," New York State Education Department, Bureau of Bilingual Education, April 1988, pp. 17-18.
24. While in 1987, the average scores for New York State females on the verbal portion of the SAT was 22 points lower than the average score for males, the average score on the math portion was 48 points lower.
25. "Scholarship Quotas," *The Times Union*, Albany, NY, May 16, 1988.
26. For example, see Patty Flakus-Mosqueda, "Critical Thinking in American Schools: Results of a Survey of Selected Teachers," *Education Commission of the States*, Working Paper No. HL-88-2, May 1988.
27. The State Education Department and a recent Comptroller's report differ in their assessment of extent of error in the grading of state exams.
28. Michael Cohen, "Designing State Assessment Systems," *Phi Delta Kappan*, April 1988, p. 584.
29. *Measuring Up: State Roles in Educational Accountability*, Policy Brief of the OERI State Accountability Study Group.

30. Part 100 of the Regulations of the Commissioner of Education, New York State Education Department, Section 100.2, Subsections m1 and m3.
31. Memo from the Commissioner of Education to the Board of Regents regarding the Regents School Improvement and Accountability Program, September 12, 1988, pp. 2-3.
32. Craig E. Richards, "Indicators and Three Types of Educational Monitoring Systems: Implications for Design," *Phi Delta Kappan*, March 1988, pp. 495-98.
33. *Accountability Reporting in the States: Report of a Survey, 1987* (working draft), State Education Assessment Center, Council of Chief State School Officers, January 1988, p. 12.
34. Richards, *op. cit.*, p. 497.
35. James R. Masters, *1985-86 Thinking Skills Testing in Pennsylvania's Student Assessment Program*. Presentation at the ECS/CDE Assessment Conference, Boulder, CO, June 1986.
36. Doug Archbald and Fred Newmann, "Beyond Standardized Testing: Assessing Authentic Achievement in the Secondary School," *National Association of Secondary School Principals*, 1988, pp. 37-39.
37. The Hawkins-Stafford Education Improvement Act of 1988, P.L. 297, as cited in "Fewer Than Half of States Join NAEP Pre-Test," *Education Week*, September 7, 1988, p. 31.
38. Lala Carr Steelman and Brian Powell, "Appraising the Implications of the SAT for Education Policy," *Phi Delta Kappan*, May 1985, pp. 604-5.
39. *Educational Achievement: Explanations and Implications of Recent Trends*, Congress of the United States, Congressional Budget Office, August 1987, pp. ix-xvii.

Bibliography

"Accountability and Governance in Education: Principles for State Legislatures." Report of the National Conference of State Legislatures Task Force on Education, January 1988.

"Accountability Reporting in the States: Report of a Survey, 1987" (working draft). Council of Chief State School Officers. State Education Assessment Center, January 1988.

Archibald, Doug and Newmann, Fred. *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School.* National Association of Secondary School Principals, 1988.

Barbieri, Edmund. "Talents Unlimited: One School's Success Story." *Educational Leadership*, April 1988, p. 35.

Barell, John; Liebmann, Rosemarie; and Sigel, Irving. "Fostering Thoughtful Self-Direction in Students." *Educational Leadership*, April 1988, pp. 14–17.

Barnes, Ronald; Moriarty, Karen; and Murphy, John. "Reporting Testing Results: The Missing Key in Most Testing Programs." *NASSP Bulletin*, November 1982, pp. 14–20.

Carelli, Frank and Lemke, William. "Exploring the Relationship Between Expenditures and State SAT Performance." A paper presented at the 1988 American Educational Finance Association Annual Conference, March 18, 1988.

Carpenter, Beth. "Translate Test Score Hieroglyphics into Clear Support for Your Schools." *The American School Board Journal*, February 1983, pp. 33, 44.

Cohen, Michael. "Designing State Assessment Systems." *Phi Delta Kappan*, April 1988, p. 584.

_____. *Restructuring the Education System: Agenda for the '90s.* National Governor's Association Center for Policy Research, December 1987.

The College Board. "1987 Profile of SAT and Achievement Test Takers." 1987.

Congress of the United States. Congressional Budget Office, August 1987, pp. ix–xvii.

Culyer, Richard C. III. "Interpreting Achievement Test Data: Some Areas of Concern." *The Clearinghouse*, April 1982, pp. 374–380.

David, Jane L. *Improving Education with Locally Developed Indicators.* The Rand Corporation Center for Policy Research in Education, October 1987.

_____. "The Use of Indicators By School Districts: Aid or Threat to Improvement?" *Phi Delta Kappan*, March 1988, pp. 499–502.

Educational Testing Service. *Testing, Equality, and Handicapped People*, 1988.

_____. "Test Use and Validity." Princeton, New Jersey, February 1980.

"Fewer than Half of States Join NAEP Pre-Test." *Education Week*, September 7, 1988, p. 31.

Flakus-Mosqueda, Patty. "Critical Thinking in American Schools: Results of a Survey of Selected Teachers." Education Commission of the States, Working Paper No. HL-88-2, May 1988.

Henson, Kenneth T. and Saterfiel, Thomas H. "State Mandated Accountability Programs: Are They Educationally Sound?" *NASSP Bulletin*, January 1985, pp. 23–27.

Jorgensen, Margaret. "Basic Differences Between Norm-Referenced and Criterion-Referenced Tests." Southern Regional Education Board, 1986.

Koretz, Daniel. "Arriving in Lake Wobegon: Are Standardized Test Exaggerating Achievement and Distorting Instruction?" *American Educator*, Summer 1988, p. 11.

Madaus, George. "Testing and the Curriculum: From Compliant Servant to Dictatorial Master." In Tanner. *Critical Issues in the Curriculum*, National Society for the Study of Education Yearbook, 1988.

_____. "Test Scores as Administrative Mechanisms in Educational Policy." *Phi Delta Kappan*, May 1985, pp. 611–617.

Masters, James R. "1985–86 Thinking Skills Testing in Pennsylvania's Student Assessment Program." Presentation at the ECS/CDE Assessment Conference, Boulder, Colorado, June 1986.

McTighe, Jay and Lyman, Frank, Jr. "Cueing Thinking in the Classroom: The Promise of Theory-Embedded Tools." *Educational Leadership*, April 1988, pp. 18–24.

New York State Education Department. Bureau of Bilingual Education. *Report on Services to Limited English Proficient Students 1984–1988*, 1988.

_____. Division of Educational Testing. "Student Achievement in New York State 1986–87."

_____. Office of District Superintendents, School District Organization and Development. "New York State School Improvement Program Guidebook." September 1987.

_____. Office for Education of Children with Handicapping Conditions. *Alternative Testing Techniques for Students with Handicapping Conditions*, May 1986.

_____. Part 100 of the Regulations of the Commissioner of Education, November 1984.

Oakes, Jeannie. *Educational Indicators: A Guide for Policymakers*. The Rand Corporation Center for Policy Research in Education, October, 1986.

Porter, Andrew. "Indicators: Objective Data or Political Tool?" *Phi Delta Kappan*, March 1988, pp. 503–508.

Ravitch, Diane. "The Uses and Misuses of Tests." *The College Board Review*, Winter 1983–84, pp. 23–26.

Richards, Craig. "Indicators and Three Types of Educational Monitoring Systems: Implications for Design." *Phi Delta Kappan*, March 1988, pp. 495–99.

"Scholarship Quotas." *The Times Union*, Albany, New York, May 16, 1988.

Smith, Marshall S. "Educational Indicators." *Phi Delta Kappan*, March 1988, pp. 487–491.

Steelman, LalaCarr and Powell, Brian. "Appraising Implications of the SAT for Education Policy." *Phi Delta Kappan*, May 1985, pp. 604–5.

U. S. Department of Education. Office of Educational Research and Improvement. "Measuring Up: State Roles in Educational Accountability." Policy Brief of the OERI State Accountability Study Group.

_____. Office of Planning, Budget, and Evaluation. *State Education Statistics: Student Performance and Resource Inputs*, February 1988.

Wolf, James M. and Kessler, Anna L. "Entrance to Kindergarten: What is the Best Age?" Educational Research Service Monograph.

Other Association Position Papers

Essential Leadership

Staff Development: Catalyst for Change

Selecting the Superintendent

Education for the Gifted and Talented

Textbook Selection: A Matter of Local Choice

A Kaleidoscope of Student Needs: New Challenges for Pupil Support Services

Meeting in the Middle: Directions in Schooling for Young Adolescents

Home-School Partnership: School Boards and Parents

The Right Start: Promises and Problems in Early Childhood Education

The Impact of Class Size on Teaching and Learning

The Vocational Mission of the Public Schools

Teacher Quality: Viewpoints on Teacher Preparation

Toward Better Teaching

Staying in School: the Dropout Challenge

Global Education and Second Language Study in the Public Schools

The Case Against Corporal Punishment